Task-Specific Cameras

Orr Zohar

Abstract—**Real world scenes have a dynamic range much larger than today's imaging sensors, leading to frequent over/under exposure of different image portions. Object detection under such extreme lighting conditions is easily confounded, which challenges existing object detection pipelines [1]. The conventional cameras' limited dynamic range stems from the global shutter and analog to digital conversion at the sensor plane, which could be thought of as a data transfer bottleneck. We can therefore formulate the problem of object detection on real-world scenes as an encoder-bottleneck-decoder scheme. Programmable photosensors – which are sensors that can perform some computation in the sensor plane itself – lend themselves as a possible solution to this problem. By developing a differentiable model for these sensors, we aim to integrate them into a pipeline that will allow us to perform end-to-end optimization of both hardware and software in unison. In this publication, we aim to introduce a *task-specific* data capture pipeline where both the hardware (focal-plane sensor-processors) and software (neural networks) are jointly and specifically optimized for one task - in this case, object detection.**

I. Introduction/Motivation

In recent years, we have witnessed the rapid acceleration of Machine Learning, producing endless novel applications. One of the more significant advancement has been in object detection/segmentation - where current state of the art (SOTA) neural algorithms are reaching unprecedented accuracy, even surpassing their human counterparts. Such advancements are extremely important for a bevy of exciting applications such as autonomous driving, personal robotics, and even in radiology. Algorithms such as YOLO [7], Mask R-CNN [9], RetinaNet [11] and others leverage deep neural architectures with sophisticated layers and losses in order to achieve state-of-the-art results. However, a key component of their success is owed to the quality of the data they are trained on. Most of them rely on datasets, such as MS-COCO [8], which contains over 1.5 million annotated object instances. However, these datasets are composed of LDR images, which cannot fully depict real-world scenes. Therefore, a key component of the success of such pipelines is the image capture itself.

Most modern digital cameras utilize the same optical design as their analog predecessors - where optics are used to create an in-focus image of the desired scene on the photosensor array within, and the array is exposed for a fixed (and constant) time interval. During the exposure, the sensors integrate the incoming luminescence, ultimately reporting the total accumulated intensity measured at each position. However, due to the limited range of intensity values these sensors are sensitive to, only a range of incoming luminescence can be accurately reported, with some values getting saturated in bright regions while others measure values below the sensor SNR (under-exposure). These effects are very detrimental to the aforementioned SOTA object detection algorithms and their deployment in real-world applications.

Different computational photography approaches have managed to capture High Dynamic Range (HDR) images - images who's dynamic range more closely approximates that of the real-world scenes - however each time there was an inherit trade-off (see related works). All of these methods are specifically optimized for visual perception, rather than any single down-stream task, such as object detection. It is fair to assume that for object detection segmentation, different details affect the accuracy of the approach. For example, edges are particularly important for object detection and segmentation, and therefore it would make sense to want to preserve this information more accurately. A new trend in Computation imaging attempts to integrate hardware and software together into a singular "neural network" utilizing differential system modeling and end-to-end optimization [4] [5] [6]. In this approach, the pipeline can be conceptualized as a neural auto-encoder, with the hardware essentially encoding the scene, followed by a second neural network that decodes the measurements. Herein, we purpose task-specific optimization using focal-plane sensor processors, allowing the development of an image acquisition pipeline optimized specifically for object detection. Here, the "camera" no longer will capture "images" in the traditional sense, but rather measurements containing the most pertinent information for object detection.

II. Related Works

HDR Imaging: The conventional camera's limited dynamic range has been extended using a variety of approaches. For example several low dynamic range images can be captured in quick succession before being combined together to create a single HDR image [12] [13]. However, this approach tends to quickly degrade in dynamic scenes, where the motion of objects cause "ghosting artifacts". Even more recently, several single-capture methods have been proposed [14]. Such approaches use a variety of ways (ND filter, SLMs, and more) to effectively have a spatially-distributed exposure time. A post-capture HDR image reconstruction step is therefore required for such approaches, and there is an inherit resolution tradeoff. Furthermore, such systems tend to be expensive and require the sensor to be permanently altered, further hindering development and prototyping. Finally, there are methods that attempt to compress the dynamic range of the signal before the bottleneck [2] [3]. Common among these approaches is the encoding of the high dynamic range information before the bottleneck (sensor dynamic range A2D), and utilizing more complex post-processing scenes to "decode" the captured image into an HDR image acquisition. However, these

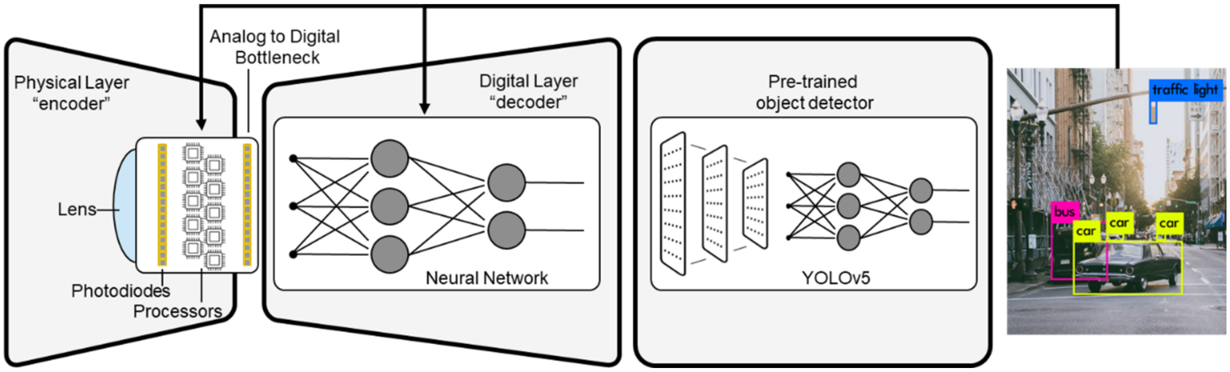


Fig. 1. Block diagram of the proposed method. An incident scene is focused on the focal-plane sensor-processors, where it is encoded ”physical encoding” before being passed through the bottleneck to the digital decoder (NN). Finally, the decoded image is passed to a frozen, pre-trained NN and the mAP loss is back-propagated to the relevant modules.

methods are rarely stable, and require extensive reconstruction schemes, limiting their frame rate and practicality.

Tone Mapping: Tone Mapping Operators (TMOs) take HDR images and map them into visually-Representative LDR images. Tone-mapping methods fall into two categories: global tone mappers and local tone mappers. Global tone mappers apply the same compression function to all pixels in the image (e.g. gamma correction). Meanwhile, local tone mappers apply pixel tone-mapping based on their neighboring pixels. Although global tone mappers are computationally more efficient, they do not preserve enough contrast, resulting in a washed-out image. Local tone mappers, on the other hand, are able to preserve contrast ratios while also adhering to regional details. An example of such local TMOs is the Reinhard tonemapper [10].

End-to-End Optimization of camera hardware and software: While the co-design of hardware and software lies at the roots of computational imaging, only recently have there been attempts to conjointly learn both the software and hardware in unison [4] [5] [6]. New tools, such as differential modeling of hardware and more advanced Neural Network tools have enabled the conceptualization of the camera as a neural auto encoder see figure 1. Here, the hardware essentially ”encodes” the incident scene before passing the measurements through the bottleneck to the neural decoder. In this research, we aim to append this traditional framework with a pre-trained YOLOv5s network, and use it to generate the ”object detection” loss.

III. Methods/Project Overview

In this project, I will attempt to implement different encoder-decoder pairs (see table I), along with a configurable ”bottleneck” quantization/saturation unit (utilizing Surrogate gradients) and perform end-to-end optimization on:

1. Identity (output = input, L1 loss)
2. Reinhard (output = Reinhard(input), L1 loss)
3. Object Detection using YOLOv5s [7]

This stepped approach will allow better initialization of our model before heading to the more complex object detection loss. For object detection, I will essentially append our model with a pre-trained neural network - YOLOv5s [7]. YOLOv5s is a state-of-the-art object detection model that takes in LDR images and outputs bounding boxes predictions and object classification probabilities. We will use the mean average precision (mAP), which is the average over multiple Intersection over Union (IoU) values for correct bounding box predictions as the loss for training. During training, we will freeze the YOLOv5s weights, thus only updating the weights of our model (encoder -decoder, **bold** in table I). Ultimately, our model will generate an LDR image given an HDR image, optimized especially for object detection. In contrast to most previously reported works, we won’t attempt to perform tone-mapping to produce visually coherent images, *but instead, produce an image optimized for object detection*. After training, for inference, an HDR image is passed to our model which outputs an image to YOLOv5s for object detection. Our end-to-end approach can be seen in figure 1.

IV. Timeline and Intermediate Goals

Week 1: I will attempt to implement 1-3 in table I without quantization.

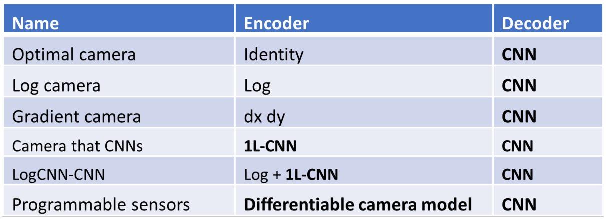
Week 2: Then, I will introduce quantization to S1.

Week 3: Once the learned decoder approaches have been realized, I will move on to the dual-learned encoder-decoders (rows 4-6), without quantization.

Week 4: I will introduce quantization to S3.

Week 5: I will develop a differentable model for the focal plane sensor-processors, implement it, and perform end-to-end optimization with/without quantization.

TABLE I
Different encoder-decoder pairs, Bold=learned



Name	Encoder	Decoder
Optimal camera	Identity	**CNN**
Log camera	Log	**CNN**
Gradient camera	dx dy	**CNN**
Camera that CNNs	**1L-CNN**	**CNN**
LogCNN-CNN	Log + **1L-CNN**	**CNN**
Programmable sensors	**Differentiable camera model**	**CNN**

References

[1] E. Onzon, F. Mannan and F. Heide, ”Neural Auto-Exposure for High-Dynamic Range Object Detection,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7706-7716, doi: 10.1109/CVPR46437.2021.00762.

[2] J. Tumblin, A. Agrawal and R. Raskar, ”Why I want a gradient camera,” 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, pp. 103-110 vol. 1, doi: 10.1109/CVPR.2005.374

[3] H. Zhao, B. Shi, C. Fernandez-Cull, S. -K. Yeung and R. Raskar, ”Unbounded High Dynamic Range Photography Using a Modulo Camera,” 2015 IEEE International Conference on Computational Photography (ICCP), 2015, pp. 1-10, doi: 10.1109/ICCPHOT.2015.7168378.

[4] J. N. P. Martel, L. K. Müller, S. J. Carey, P. Dudek and G. Wetzstein, ”Neural Sensors: Learning Pixel Exposures for HDR Imaging and Video Compressive Sensing With Programmable Sensors,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 7, pp. 1642-1653, 1 July 2020, doi: 10.1109/TPAMI.2020.2986944.

[5] Chang, J., Sitzmann, V., Dun, X. et al. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Sci Rep 8, 12324 (2018). https://doi.org/10.1038/s41598-018-30619-y

[6] Metzler, C., Ikoma, H., Peng, Y., Wetzstein, G., Deep Optics for Single-shot High-dynamic-range Imaging, CVPR 2020

[7] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. ”You only look once: Unified, real-time object detection.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.

[8] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. ”Microsoft coco: Common objects in context.” In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

[9] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. ”Mask r-cnn.” In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.

[10] E. Reinhard and K. Devlin, ”Dynamic range reduction inspired by photoreceptor physiology,” in IEEE Transactions on Visualization and Computer Graphics, vol. 11, no. 1, pp. 13-24, Jan.-Feb. 2005, doi: 10.1109/TVCG.2005.9.

[11] Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. ”Focal loss for dense object detection.” In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988. 2017.

[12] P. E. Debevec and J. Malik, “Recovering high dynamic range radiance maps from photographs,” in Proceedings of the 24th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 1997, pp. 369–378.

[13] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, “Burst photography for high dynamic range and low-light imaging on mobile cameras,” ACM Transactions on Graphics (TOG), vol. 35, no. 6, p. 192, 2016.

[14] S. K. Nayar and T. Mitsunaga, “High dynamic range imaging: Spatially varying pixel exposures,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2000, pp. 472–479.